CORE
Metadata, citation and similar papers at core.ac.uk
Provided by Kent Academic Repository

Citation:

Brown, A. (in press). Item Response Theory approaches to test scoring and evaluating the score accuracy. In Irwing, P., Booth, T. & Hughes, D. (Eds.), The Wiley Handbook of Psychometric Testing. London: John Wiley & Sons.

Chapter 21

Item Response Theory approaches to test scoring and evaluating the score accuracy

Anna Brown

University of Kent

Abstract

The ultimate goal of psychometric testing is to produce a score by which people can be differentiated. Item Response Theory (IRT) devises methods for estimating person's score on one or more psychological constructs (traits) from his/her responses to test items. This chapter gives an overview of scoring methods applicable to situations when the test items indicate one trait only; or a set of related traits but each item contributes to measurement of one trait; or when each item indicates multiple traits. We consider scoring methods based on item responses only, as well as Bayesian methods, which use prior knowledge of the trait distribution. Much of this chapter is devoted to methods for assessing measurement precision provided by individual items, the whole test, and the prior distribution. In IRT, this precision can be evaluated for each individual response pattern. All described methods are illustrated with a single empirical example.

Keywords: test scoring; IRT person parameter; item information function; prior information; posterior information; multidimensional IRT; bifactor model; marginal reliability.

The ultimate goal of psychological measurement is to produce a score by which people can be assessed and differentiated. Item Response Theory (IRT) views test items as a series of small experiments, "from which a measure is inferred" (van der Linden and Hambleton 1997). In IRT, responses to test items serve as indicators of a person's standing on some underlying psychological construct or constructs, and devises special algorithms for determining that standing. The purpose of this chapter is to give an overview of IRT methods for inferring person's scores on the psychological constructs of interest, often referred to as *abilities*, *proficiencies* or *traits* (we will call them *traits*).

To be useful in applications, the score must infer the person standing on the trait continuum accurately, and importantly, the precision level must be known for decision-making purposes. IRT has many advantages over Classical Test Theory (CTT) in estimating both the score and its precision. With IRT, we can control for properties of test items – such as difficulty or liability to guessing – making the score independent of these nuisance factors. With IRT, we can also drop unattainable assumptions of continuity for dichotomous test items (correct-incorrect, or yes-no), and of equal intervals in rating categories (for example, 5-point scales with response options ranging from "strongly disagree" to "strongly agree"). Treating categorical item responses appropriately brings the test scores much closer to the interval level of measurement so that the standard statistics can be applied to them. With IRT, we can also drop an unattainable assumption that the precision of test scores is a single value that holds for a sample, and assess the measurement precision for each individual response pattern. In many testing contexts, knowing the measurement precision associated with a particular pattern (and score) enables better judgments about significance of difference between any two respondents, or any change occurring in scores, for instance in response to treatment etc. (Reise and Haviland 2005). At the same time, we often need to summarize the

overall precision of measurement in a research sample, or in the population as a whole – and IRT provides methods for that too.

In this chapter, we attempt to make these methods more readily available to students and researchers by providing formulae for scoring and precision estimation suitable for most commonly used models – a single factor model, a correlated factor model, and a bifactor model. All described methods are illustrated with a single data analysis example involving a short patient satisfaction measure, the Experience of Service Questionnaire (ESQ), completed by parents of children treated for mental health problems.

The multidimensional item response model

Psychometric tests often necessitate the capture of several related constructs. For example, several cognitive facets, which are correlated with each other, might be of interest. In mental health measures, we might be interested in capturing several distinct areas of functioning, which might also form an overall domain. To devise scoring methods suitable for all such measures, multidimensional factor models are recommended (Gibbons, Immekus & Bock 2007). In this section, we provide a brief overview of the core concepts of IRT and some general references necessary for this chapter. For more detailed introduction, see chapter 17.

Let $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_T)$ (pronounced '*theta*') be a set of T unobserved, or *latent* traits (we may also call them abilities, proficiencies, constructs or dimensions) measured by a psychometric test. In the simplest case $T = 1$, and we deal with a test measuring just one trait. Such models are generally referred to as *unidimensional* models (See Chapter 16). In all other cases, $T > 1$ and models are referred to as *multidimensional*. In both cases, the latent traits are assumed normally distributed, have mean zero and unit variance. In multidimensional

models, latent traits may also be correlated, with their covariance matrix denoted Σ (pronounced *'sigma'*).

Each test item is designed to measure one or more traits (usually one, but we will see that cases when an item contributes to measurement of two traits are quite common). An item elicits an internal response from a participant. To describe this internal item response, we use the notion of *response tendency*. The **unobserved** response tendency is most likely a complex process within individuals, which we cannot access or measure directly. Instead, we assess an **observed** item response u_i. Observed item responses are sometimes dichotomous ('yes' – 'no' , 'agree' – 'disagree'), and often polytomous ('never' – 'sometimes' – 'often' – 'always'; or 'strongly disagree' – 'disagree' – 'neutral' – 'agree' – 'strongly agree' etc.). From dichotomous responses, we do not know the exact extent of agreement with a statement; all we know is that the respondent picked 'agree' out of two available response options. Although the polytomous options provide more opportunities to quantify the extent of agreement, the exact level of internal agreement (the response tendency) is not observed – only its categorization into one of response options is observed.

The observed response u_i relates to the unobserved response tendency u_i^* through a threshold process. There is one threshold when two response alternatives are used. When the response tendency is above the threshold, the keyed response $u_i = 1$ is given; and when the response tendency is below the threshold, the non-keyed response $u_i = 0$ is given. In addition, we assume that the unobserved response tendency u_i^* is caused by one or more traits and can be described by a linear factor model of Spearman (the response tendency is a linear function of one or more thetas). For example, the tendency to solve a problem on a mathematics test increases as the mathematical ability increases; and if the solution achieved for this problem

is enough to provide the answer (the response tendency is above the threshold), the correct response is given.

For ease of exposition, in what follows we give formulae for the dichotomous case, mentioning how to extend them to polytomous cases in passing. Polytomous models can be achieved by considering the probability of choosing (or otherwise) each response option, as is done in, for example, a graded response approach (Samejima 1969), or a partial credit model (Masters and Wright 1997). Not one but several thresholds are considered in this case (k –1 where k is the number of options), each representing a boundary between selecting one of the two adjacent response categories.

Dichotomous events such as passing / failing a test item are commonly described in terms of their probability, which directly depends on the response tendency – and consequently on the latent trait or traits. The probability of passing a mathematics item, for example, increases as the ability increases. This increase is not linear but s-shaped, with slow increases of the probability of passing at the extremes of ability scale, and more rapid increase in the range of the item threshold. A well-known function with the needed shape that is commonly used to describe the link between the probability and the response tendency is the cumulative standard normal distribution function (aka *normal ogive*).

With this, the *item response function* (IRF) for item i measuring T traits is given by

$$P_i = P\left(\mathrm{u}_i = 1 \middle| \boldsymbol{\theta}\right) = \Phi\left(-\tau_i + \beta_{i1}\theta_1 + \ldots + \beta_{iT}\theta_T\right), \qquad (0.1)$$

where τ_i is the item *threshold*, and β_{ki} is the *slope* for k-th trait – describing how fast the probability of the keyed response changes with the unit change in trait k^{i}. $\Phi(x)$ denotes the cumulative standard normal distribution function evaluated at x. Without loss of generality, we use the normal-ogive (or *probabilistic*) link function here. Alternatively, the *logistic* link function $\mathrm{L}\left(x\right) = 1/\left(1+e^{-x}\right)$ can be used (Reckase 2009).

This general model allows items to measure any combination of *T* traits by having non-zero slopes on some traits and zero on others. Thus, we can easily accommodate the simplest situation when each test item indicates only one trait – the model possesses an *independent-clusters* structure (McDonald 1999) – but the test overall may measure more than one trait.

Latent trait estimation

In Item Response Theory, the latent trait scores $\boldsymbol{\theta}$ can be estimated by treating the model parameters (item threshold and slopes, and the correlations between traits) as if they were known. This is reasonable if model parameters have been accurately estimated. When the item and other model parameters are known, the IRF depends only on the latent traits, and the fundamental approach to estimating the trait scores is to search for values that maximize the likelihood of the observed pattern of responses $\mathbf{u} = (u_1, u_2, ..., u_m)$ to all *m* items in the test. To proceed with the estimation, we make an assumption of *local independence,* which states that in a subpopulation where the latent traits take fixed values (a subpopulation of people who have the same latent trait scores) the item responses are independent.

Given that the item responses are independent (conditional on the latent traits), it is easy to express the probability of the observed pattern as the product of probabilities of the responses to individual items. For items to which the keyed response ($u_i = 1$) was given, the probability P_i is given by (0.1). For items to which the non-keyed response ($u_i = 0$) was given, the probability is $Q_i = 1 - P_i$. The *maximum likelihood* (ML) scores are found iteratively by searching for a set of trait scores $\boldsymbol{\theta}$ that maximize the *likelihood function* – the product of the probabilities of all given responses:

$$l(u_1, u_2, ..., u_m | \boldsymbol{\theta}) = \prod_{u_i=1} P_i(\boldsymbol{\theta}) \prod_{u_i=0} (1 - P_i(\boldsymbol{\theta})). \qquad (0.2)$$

Maximum likelihood scores only use information contained in the item responses, and therefore are philosophically uncontroversial (McDonald 2011). They, however, are undefined for some response patterns, notably for "perfect" patterns when the respondent gave keyed responses to all items (for instance by answering "yes" or selecting the top rating category such as "strongly agree"). This situation is illustrated in Figure 1a, where probabilities of observed responses to two test items conditional on the latent trait are shown, as well as the joint likelihood of these responses. In this case, the maximum likelihood estimate does not exist because the joint likelihood increases infinitely when the latent trait score increases[ii]. The score is also undefined when non-keyed responses are given to all items (for instance by answering "no" or selecting "strongly disagree").

INSERT FIGURE 1 ABOUT HERE

To avoid this indeterminacy and improve estimation efficiency, prior information about the trait score distribution may be used in addition to the observed item responses. The basis for incorporating this information is given by the Bayes theorem, which, applied to the scoring problem, suggests that the probability of observing a particular ability level (theta score) given the observed response pattern is the product of two probabilities: the probability of observing the assumed ability, and the probability of observing the response pattern given the ability. More formally, in a Bayesian approach, the likelihood of trait scores θ given the observed response pattern (*posterior* likelihood l_P) is computed by multiplying: 1) the likelihood of the observed response pattern given the trait scores θ, and 2) the likelihood of the theta scores occurring in the population (*prior* distribution; usually multivariate standard

normal). The former is, of course, the likelihood (0.2) used in ML estimation, and the latter is the normal density function ϕ, thus

$$l_P\left(\mathrm{u}_1,\mathrm{u}_2,...,\mathrm{u}_m\middle|\boldsymbol{\theta}\right)=\phi(\boldsymbol{\theta})\prod_{\mathrm{u}_i=1}P_i(\boldsymbol{\theta})\prod_{\mathrm{u}_i=0}\left(1-P_i(\boldsymbol{\theta})\right). \tag{0.3}$$

An example posterior likelihood function is given in Figure 1b, where IRF for two test items are shown together with the normal density function, and the joint likelihood is the product of all three functions. In this case, the maximum likelihood estimate exists because the posterior likelihood function has a single peak. By adding information from the assumed multivariate normal distribution of scores, the problem of undefined trait scores for perfect patterns is overcome, that is, a score is always defined when the Bayesian estimation is used.

Two computational methods for test scoring using the Bayesian approach are a) *expected a posteriori* (EAP) estimation, which computes **the mean** of the posterior distribution of the likelihood; and b) *maximum a posteriori* (MAP) estimation, which computes **the mode** of the posterior distribution (Embretson and Reise 2000). The EAP method is an excellent computational option for one-dimensional tests. The mean of the posterior likelihood is approximated taking "snapshots" of the continuous likelihood function at *q* points (*quadrature* points) selected along the latent trait continuum. Formula for computing the EAP scores for a single trait is given in Appendix A. In this formula, each quadrature point value θ_q is weighted by the value of the posterior likelihood function at that point, the weighted sum of all q points is computed, and then divided by the sum of weights. For this approximation of the mean value of the continuous distribution of the likelihood to be accurate, the choice of the quadrature points is important. The quadrature points are usually set at equal intervals, and a larger number of points would yield a more precise mean value.

For a discussion regarding the number of points necessary for precise estimation, see Thissen and Orlando (2001).

The EAP score is easy to compute when only one dimension is involved. EAP estimation, however, becomes computationally demanding as the number of traits increases. This is because with two traits, one needs to sample q^2 quadrature points, for every combination of the theta values for trait 1 and trait 2. One needs to create a multidimensional grid of q^T points to compute an EAP score on T traits. Even with a small number of points on each trait continuum such as q =11, the resulting number of points for 2 traits is manageable q^2 = 121; for 5 traits it is already q^5 =161,051, and the number of points for 10 traits is almost 26 billion. Clearly, a less computationally demanding approach is needed in this case.

The MAP score corresponds to the mode of the posterior distribution. Finding a set of T trait scores that maximize the multidimensional posterior likelihood function requires iterative procedures using gradients. When T = 1 and only one trait is measured, this may be an unnecessary complication; however, in multidimensional models this estimation is much quicker than the EAP because it searches for the optimal set of theta values for all traits simultaneously and the number of iterations is not affected by the number of traits.

When an appropriate prior distribution is used, the Bayesian EAP and MAP approaches have been shown to achieve accurate estimates of the latent trait with fewer items than the ML method; however, they are known to shrink the latent trait distribution towards the population mean. In practice, this yields estimated scores with smaller variance than was assumed for the latent traits. The amount of shrinkage depends on several factors, including the test length, and can be quite substantial when the number of items in the test is small (Thissen and Orlando 2001). Another concern with the use of Bayesian estimation methods is that when multivariate priors are used, correlated but conceptually distinct traits will influence each other's score estimates, or "borrow strength" from each other. Some authors,

notably McDonald (2011, 535) argued against this inadvertent use of information on both philosophical and statistical grounds. Indeed, the fact that a person's ability in English may be judged by his/her results in mathematics may seem unjustified or even unfair. There are also measurement-related concerns with the use of multivariate priors, which we will discuss in due course.

Standard Error of measurement, test information and reliability

The IRT scoring methods are only our best guess at estimating the true scores for people taking tests, and inevitably, all estimation methods are associated with a certain degree of error. The joint likelihood of the response pattern and the posterior likelihood functions describe probability values for a whole range of trait scores. These distributions are typically Gaussian in appearance and have a single peak (for example, see Figure 1b). The mode (or the mean) of these distributions provides a limited summary of the likelihood function. The width or the spread of the likelihood, on the other hand, indicates the degree of uncertainty around the score estimation – the narrower the spread, the more confident we are that the true theta value is in close range of the estimated value. Responses that are less likely given the trait score, for instance an incorrect response to an easy question when ability is high, or a correct response to a difficult question when ability is low (so-called *aberrant* responses), will make the spread of likelihood values around the estimated theta wider. Responses that are in line with the estimated trait score will make the spread of likelihood values around the estimated theta narrower.

For approximately Gaussian distributions, the spread of the likelihood values is meaningfully described by the distribution's standard deviation. The standard deviation of the likelihood of the response pattern in ML estimation, or standard deviation of the posterior distribution in Bayesian estimation, therefore, is a measure of the **Standard Error** (SE) of measurement in IRT.

According to the estimation methods used, there are two main ways of computing the standard error of estimation. For the EAP estimator, it is natural to compute the standard deviation from the mean (which is the estimated EAP score) of the likelihood values taken at the quadrature points. Therefore, the standard error of the EAP estimator is based on direct evaluation of the standard deviation of the posterior distribution (the computational formula is given in Appendix A).

For the methods maximizing the mode of the likelihood function (ML and MAP), the variance of the likelihood function along the trait continuum θ is computed as the inverse of the *Fisher information* $\mathcal{I}$ (or simply *information*). The more information an item or a set of items provide for measuring latent traits, the more accurate the score estimation, and consequently, the smaller the standard error will be. For the **one-dimensional** case, the standard error of the estimated ML score $\hat{\theta}$ is

$$SE_{ML}\left(\hat{\theta}\right) = \frac{1}{\sqrt{\mathcal{I}\left(\hat{\theta}\right)}}. \tag{0.4}$$

The standard error of the MAP score involves posterior information I_p (information provided by the test items together with the prior distribution) instead:

$$SE_{MAP}\left(\hat{\theta}\right) = \frac{1}{\sqrt{\mathcal{I}_P\left(\hat{\theta}\right)}}. \tag{0.5}$$

In the **multidimensional** case, the standard error of the estimated ML score $\hat{\theta}_a$ involves computing the information in the direction of trait θ_a evaluated at the point-estimates $\boldsymbol{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_T)$

$$SE^a\left(\boldsymbol{\theta}\right) = \frac{1}{\sqrt{\mathcal{I}^a\left(\boldsymbol{\theta}\right)}}, \tag{0.6}$$

and the standard error of the MAP score $\hat{\theta}_a$ involves computing the posterior information in the direction of trait θ_a

$$SE^{a}_{MAP}\left(\boldsymbol{\theta}\right) = \frac{1}{\sqrt{\mathcal{I}^{a}_{P}\left(\boldsymbol{\theta}\right)}}. \tag{0.7}$$

The following section shows how to compute the ML and the posterior information in both one-dimensional and multidimensional cases.

Item Information Function (IIF)

When a test measures only **one trait**, the amount of information that item *i* provides toward measurement of the trait is given by the *Item Information Function* (IIF)

$$\mathcal{I}_i\left(\theta\right) = \frac{\left[P_i'\left(\theta\right)\right]^2}{P_i\left(\theta\right)\left[1-P_i\left(\theta\right)\right]}, \tag{0.8}$$

where $P_i'\left(\theta\right)$ denotes the first derivative of the item response function (McDonald 1999). Because derivatives reflect the degree of change in the probability of the keyed response with change in the trait score, the item **slope** is at the heart of the information. At the theta value corresponding to steepest slope (in binary IRT models without guessing, this point corresponds to the item difficulty), the probability of keyed response in response to the change of theta changes faster than at any other point long the trait continuum. The item discriminates best around this theta value, or, in other words, provides most information for the trait estimation. Another theta value may correspond to a shallow slope, indicating little change in the probability of the keyed response in response to change in theta.

It can be seen that, unlike in classical psychometric test theory, the precision of measurement in IRT depends on the latent trait, and therefore on item responses. This means that persons with different item responses may potentially have scores estimated to different degrees of accuracy.

When we use the normal ogive link, as in (0.1), then the expression for IIF becomes

$$\mathcal{I}_i(\theta) = \frac{\beta_i^2\left[\phi(-\tau_i + \beta_i\theta)\right]^2}{\Phi(-\tau_i + \beta_i\theta)\left[1 - \Phi(-\tau_i + \beta_i\theta)\right]}, \tag{0.9}$$

where $\Phi(x)$ denotes the cumulative standard normal distribution function; and $\phi(x)$ denotes the standard normal density function evaluated at x. The item information function is described by a curve conditional on the latent trait, and is sometimes called *item information curve*[iii].

For items with graded response categories, the item information can be derived from category response curves as follows (Samejima 1997; Dodd, DeAyala and Koch 1995):

$$\mathcal{I}_i(\theta) = \sum_{x=0}^{k} \frac{\left[P_{ix}'(\theta)\right]^2}{P_{ix}(\theta)}. \tag{0.10}$$

When items contribute to measurement of **two or more latent traits**, the *direction* of information must be considered when computing the item information function (Ackerman 2005; Reckase 2009). Let $\mathbf{d}$ be a vector of angles $\mathbf{d} = (d_1, d_2, ..., d_T)$ to all T axes that defines the direction from a point $\boldsymbol{\theta}$ in the trait space. Then the information provided by item i in direction $\mathbf{d}$ is described by a surface

$$\mathcal{I}_i^{\mathbf{d}}(\theta) = \frac{\left[\nabla_{\mathbf{d}} P_i(\boldsymbol{\theta})\right]^2}{P_i(\boldsymbol{\theta})\left[1 - P_i(\boldsymbol{\theta})\right]}, \tag{0.11}$$

where $\nabla_{\mathbf{d}} P_i(\boldsymbol{\theta})$ is the gradient in direction $\mathbf{d}$ given by (Reckase 2009):

$$\nabla_{\mathbf{d}} P_i(\boldsymbol{\theta}) = \frac{\partial P_i(\boldsymbol{\theta})}{\partial \theta_1}\cos d_1 + \frac{\partial P_i(\boldsymbol{\theta})}{\partial \theta_2}\cos d_2 + ... + \frac{\partial P_i(\boldsymbol{\theta})}{\partial \theta_T}\cos d_T. \tag{0.12}$$

Test Information Function (TIF)

When local independence holds, the Test Information Function (TIF) for a trait can be computed as the sum of the item information functions contributing to measurement of the trait. In the **one-dimensional** case, the TIF is simply

$$\mathcal{I}(\theta) = \sum_{i=1}^{m} \mathcal{I}_i(\theta). \tag{0.13}$$

In the **multidimensional** case, the total information about trait θ_a is the sum of all IIFs in direction of that trait

$$\mathcal{I}^a(\boldsymbol{\theta}) = \sum_{i=1}^{m} \mathcal{I}_i^a(\boldsymbol{\theta}). \tag{0.14}$$

Note that some items might not contribute to measurement of the focus trait, as is the case with items forming an independent clusters structure. Then, the item's slope on the trait is zero – and the information it contributes is also zero.

So far, we have only taken into account information provided by the item responses, i.e. the maximum likelihood (ML) information. However, when Bayesian estimation is used, prior information about distribution of the traits in the population will also contribute to the estimation of the latent trait. The total information from the item responses and the prior distribution is given by *posterior test information*, $\mathcal{I}_P$. For a **single trait**, a normally distributed prior with variance σ^2 adds $1/\sigma^2$ to the information uniformly across the latent trait continuum (Du Toit 2003). Since we assumed the trait variance is unity, the posterior test information is

$$\mathcal{I}_P(\theta) = \mathcal{I}(\theta) + 1. \tag{0.15}$$

For **multiple traits**, the information given by the prior distribution is added to the multidimensional ML test information given in (0.14). Assuming the multivariate standard normal prior with correlation matrix $\boldsymbol{\Sigma}$, the posterior test information is

$$\mathcal{I}_P^a(\boldsymbol{\theta}) = \mathcal{I}^a(\boldsymbol{\theta}) + \left[\boldsymbol{\Sigma}^{-1}\right]_a, \tag{0.16}$$

where $\left[\boldsymbol{\Sigma}^{-1}\right]_a$ is the a^{th} diagonal element of the inverted trait correlation matrix, $\boldsymbol{\Sigma}^{-1}$ (e.g. Brown & Maydeu-Olivares, 2011). When all traits are uncorrelated, the term added to the ML test information in equation (0.16) equals 1. When all the traits are correlated positively, the additional term is greater than 1 (and therefore the prior distribution contributes more information for the trait estimation).

Reliability

Reliability in classical test theory is defined as proportion of variance in the observed score due to the true score. This proportion is a single value within the sample on which the reliability is computed. Reliability is therefore sample-dependent, but independent of the test score in the classical account; it is the same for all people (and therefore test scores) within the same sample.

Information functions in IRT describe the precision of measurement provided by a test (and all its items) more completely than a single reliability coefficient. However, sometimes it is convenient in applications to summarize the information values into a single index. Such an index (rather than curves and surfaces) is more likely to appeal to, and be understood by, the test user, as it allows direct comparisons with classical test theory's reliability statistics. The reliability coefficient enables a quick evaluation of the test's overall measurement precision. Another important use for the reliability coefficient is predicting the relationship between the estimated and the true score, using

$$\mathrm{corr}(\theta, \hat{\theta}) = \sqrt{\rho}\,. \tag{0.17}$$

An appropriate index, *marginal reliability*, was suggested by Green and colleagues (1984):

$$\rho = \frac{\mathrm{var}[\theta] - \overline{SE^2}[\theta]}{\mathrm{var}[\theta]} = 1 - \frac{\overline{SE^2}[\theta]}{\mathrm{var}[\theta]}\,. \tag{0.18}$$

This coefficient uses the classical definition of reliability as proportion of variance in the test score due to true score. The true score variance is computed as the test score variance minus error variance (squared standard error). The reliability increases as the standard error decreases; it approaches 1 as the standard error approaches 0.

There are two ways to compute the marginal reliability coefficient. *Theoretical reliability* (Du Toit 2003) considers the **theoretical distribution** of the latent trait (which we assume standard normal), thus var[θ]=1, and formula (0.18) becomes

$$\rho = 1 - \overline{SE^2}[\theta]\,, \tag{0.19}$$

and the average squared standard error is the integral

$$\overline{SE^2}[\theta] = \int_{-\infty}^{\infty} SE^2(\theta)\phi(\theta)\,d\theta\,. \tag{0.20}$$

The squared standard errors are computed for the theoretical distribution of the latent trait from the test information function using formulae (0.4) or (0.5); therefore, the theoretical reliability coefficient is suitable for ML and MAP scores. In practice, the integral is approximated by evaluating the squared errors and the normal densities at multiple points taken at equal intervals along the trait continuum. The simple formula for theoretical reliability (0.19) allows connecting some established benchmarks for classical reliability with corresponding values of the IRT standard error and information. The reliability ρ = .75 corresponds to the squared standard error 0.25 (*SE* = 0.5), which in turn corresponds to

information $I = 4$. The reliability $\rho = .90$ corresponds to the squared standard error 0.10 ($SE = 0.32$), which in turn corresponds to information $I = 10$.

An alternative approach to obtaining marginal reliability, *empirical reliability* (Du Toit 2003), considers the standard errors not for a theoretical distribution of theta, but for the estimated sample scores.

$$\rho = 1 - \frac{\overline{SE}^2\left[\hat{\theta}\right]}{\mathrm{var}\left[\hat{\theta}\right]}. \tag{0.21}$$

Given a sample of N respondents, the trait variance is computed as the variance of the estimated theta score in the sample, and the average squared SE is computed by averaging the squared SEs of estimated theta scores $\hat{\theta}_j$ for each respondent j,

$$\overline{SE}^2\left[\hat{\theta}\right] = \frac{1}{N}\sum_{j=1}^{N} SE^2\left(\hat{\theta}_j\right). \tag{0.22}$$

The standard errors of person scores can be computed either using the standard deviation of the posterior likelihood (21.24) for the EAP score, or the inverse of Fisher information (0.5) for the ML or MAP score.

The empirical reliability is particularly quick and easy to compute when IRT software programs provide the standard errors for person scores as part of the scoring process. Since the observed score variance is known for the sample (it is simply the variance of the estimated scores), the empirical reliability can be easily computed. One simply needs to square the provided standard error values for all people, and compute the mean of the squared values to obtain the average error variance.

Applying IRT scoring methods and estimating measurement accuracy in practice

In this section, we show how person responses are scored and their standard errors are computed for a range of item response models popular in testing applications. Specifically, a

unidimensional model, a correlated traits model, and a bifactor model are illustrated with a simple data analysis example.

Data example with the Experience of Service Questionnaire (ESQ)

Questionnaire. To illustrate methods described in this chapter, we consider data from a short questionnaire measuring patient satisfaction in child healthcare. The Experience of Service Questionnaire (ESQ) was developed from focus groups with children and parents across the child health sector, identifying elements that are important for positive experience of care (Attride-Stirling 2002). Here we consider the parent version of ESQ (given in Appendix B), which is intended for use with parents/carers of young children and adolescents.

The ESQ parent version includes 12 questions about the parent's experience with service that their child received. Questions also tap into parent-centered experiences, such as whether the parent felt that *he/she* was listened to, or *his/her* problems were addressed. The version uses affirmative statements, for example "It was easy to talk to the people who have seen my child", and three response options ('certainly true'–'partly true'–'not true'). An appropriate coding of ESQ item responses is assigning consecutive integers to each response category, in accordance with the increasing level of agreement so that higher scores would represent higher levels of satisfaction. Since all questions indicate positive aspects of experiences, the appropriate coding would be **0** for the least favorable rating ('not true'), **1** for the intermediate rating ('partly true'), and **2** for the most favorable rating ('certainly true').

Sample. Our example dataset comprises responses from $N = 716$ parents of children aged between 3 and 16 years (median age 11 years), who were treated for various mental health problems in one UK member service of Child and Adolescent Mental Health Services (CAMHS). This is part of a larger multi-service sample analyzed by Brown, Ford, Deighton and Wolpert (2012).

Item endorsement rates. Distributions of responses to all ESQ items are highly skewed, with vast majority of responses falling within the category 'certainly true', which represents most favorable ratings. For most items, approximately 70-80% of all parents choose the most favorable rating. Item 3 ('treated well') shows the highest endorsement (91% of parents/carers choose 'certainly true'). The least endorsed item is that concerning appointment times (only around 59% of all parents chose 'certainly true'). Considering whole response patterns rather than responses to individual items, 28.2% of parents endorse the top rating category for all items, and only 0.1% endorse the bottom rating category for all items.

Measured constructs. In previous analyses conducted by Brown and colleagues (2012), the ESQ was shown to measure two highly correlated aspects of satisfaction, satisfaction with Care, and with Environment. We start by exploring the factor structure of these data by performing an exploratory factor analysis (EFA) with categorical variables in M*plus* (Muthén and Muthén 1998-2012), using full information maximum likelihood (FIML) estimation. Samejima's (1969) graded response model is used by Mplus to model polytomous responses.

The first four eigenvalues for this analysis are 8.13, 1.23, 0.71 and 0.56. As the ratio of the first to the second eigenvalue is very large, a strong general factor is evident, together with one further factor. Goodness of fit of the exploratory two-factor solution is significantly better than the one-factor solution (likelihood ratio test reported by Mplus $\chi^2 = 128.6$, df = 11, $p < 0.001$).

An oblique rotation of two factors yields nearly ideal independent clusters, complying with the previously reported structure. The first factor is indicated by nine items: 1, 2, 3, 4, 5, 6, 7, 11 and 12. These items relate to satisfaction with **Care** including quality of communication, competence of medical staff and consistency of care. The second factor is indicated by items 8, 9 and 10, which relate to satisfaction with **Environment** surrounding

the treatment, such as appointment times, facilities and location. Item 3 has a non-trivial cross-loading on this factor, suggesting that being "treated well" means good customer service in general as well as good medical help. As expected, the two aspects of satisfaction are moderately correlated at 0.60.

It has been argued that presence of a strong 'halo' effect is evident in responses to the ESQ (Brown et al. 2012), because even theoretically unrelated aspects of service experience (i.e. appointment times, facilities and location) correlate with each other and with care-related aspects. Global *affective satisfaction* has been suggested as the likely explanation of this halo effect, and an alternative model for item responses has been proposed whereby all item responses are underlain by the Affective Satisfaction factor, and in addition, care-related items are underlain by a specific Experience of Care factor.

To illustrate the process of estimating test scores and their precision using unidimensional and multidimensional IRT approaches, we consider three alternative conceptual measurement models for the ESQ.

The first model is a **Unidimensional** model, in which one common factor (presumably satisfaction with service) explains all variation in the data (see Figure 2a). This is the most basic IRT model, and is the crudest representation of these data. The unidimensional model assumes that the ESQ items are independent controlling for global satisfaction.

The second model is a **Correlated Traits** model, in which the two aspects of satisfaction, satisfaction with care and environment, are indicated by their respective items and correlate freely (Figure 2b). This model is useful if the focus of measurement is to differentiate between the two domains of satisfaction.

The third model is a **Bifactor** model, in which the care-related items indicate two factors – the general Affective Satisfaction factor and the specific Experience with Care

factor (see Figure 2c). The environment-related items indicate the general factor only, assuming that there is no further common reason for co-variation between these items. The general factor accounts for the common variance shared by all items, and the specific factor is the 'residual' dimension uncorrelated with the general factor, accounting for any remaining common variance specific to care-related experiences.

INSERT FIGURE 2 ABOUT HERE

Scoring under the Unidimensional IRT model

We know that two factors underlie the ESQ data rather than one; however, for purposes of illustration we proceed as if the ESQ really did measure only one underlying trait, 'satisfaction with service'. This analysis will be later compared with other analyses using other, more suitable models. Assuming a one-dimensional structure when more than one factor is present would cause local dependencies between some ESQ items. When the local independence assumption does not hold, maximum likelihood estimation might produce biased results. In addition, the standard errors might be biased because when local dependencies exist, the test information cannot be decomposed into the sum of item information functions.

Illustration: ESQ satisfaction scored by the unidimensional EAP method

We estimate item parameters according to the unidimensional model depicted in Figure 2a, using the ML estimator with the probabilistic link (LINK=PROBIT) in M*plus*. Table 1 reports the loglikelihood with the number of estimated parameters, and the Bayesian information criterion (BIC; Schwarz, 1978) for this and other models. These values, although not informative on their own, may be used for comparison of alternative models

---

INSERT TABLE 1 ABOUT HERE

Table 2 gives the slopes for all 12 items; Table 3 gives the item thresholds. It can be seen that all the thresholds are large negative values, indicating high levels of endorsement ("easiness") of all items.

INSERT TABLES 2 AND 3 ABOUT HERE

Next, we estimate the person satisfaction scores using the EAP method. The EAP scores range from -3.12 to 0.91 (mean = -0.002, SD = 0.88). The standard errors of the EAP scores range from 0.17 to 0.73. Figure 3 shows the standard errors plotted against the EAP trait scores for all parents in the sample. For the low end of satisfaction (scores ranging between –3.0 and 0.5), the precision of measurement is good, with the standard errors below 0.5. For the scores between approximately -2.5 and 0, the standard errors are 0.3 or below. We, however, know that the unidimensional model ignores the local dependencies between items (environment related items 8-10, as high residuals show). Ignoring local dependencies when they exist may lead to inflated estimates of test information (Thissen et al. 2001), which in turn leads to deflated standard errors. Thus, the obtained standard errors are probably lower than they should be.

We conclude that the scale discriminates well between parents with low to average levels of satisfaction, but lacks ability to differentiate between higher scores. The result is a profound 'ceiling effect', whereby a very large group of parents (28.2%) who gave the top rating to all experiences, received the same score (see the histogram for theta score in Figure 3). The top estimated theta score has a large standard error, because the test items cannot differentiate between these respondents.

INSERT FIGURE 3 ABOUT HERE

It can be also seen from Figure 3 that the vast majority of points on the graph can be approximated by a curve. This curve is the standard error function computed from the posterior test information function using formula (0.5). Because the item and test information functions depend on the theta but do not depend on the observed response pattern, the standard error function yields the same value for all respondents with the same estimated theta score. This is not true for the EAP-estimated standard errors, which depend not only on the estimated theta but also on the observed response pattern, as can be seen in formula (21.24). Therefore, some EAP standard errors on Figure 3 deviate somewhat from the smooth SE function. The standard errors are typically larger for those with aberrant response patterns – for instance, those parents who agreed with one item but disagreed with another, similar item. We compute empirical reliability of the EAP scores by averaging the squared standard errors of observed scores, which are produced and saved by Mplus for every respondent, and by obtaining the observed score variance from Mplus output. The squared values of standard errors average at 0.229, the variance of the observed score is 0.773, and the empirical reliability is therefore $\rho = 1 - (0.229/0.773) =$ **0.70**.

To give some comparison with commonly used classical reliability statistics, we compute Cronbach's alpha, which is 0.90 for this scale (see Table 4 for comparison with other models). This example is a good illustration of pitfalls of using a single summary coefficient to describe the complexity of standard errors conditional on response patterns. We know that measurement accuracy is different in different ranges of the trait; the standard error functions describe that accuracy fully, while the reliability indices merely give an aggregated picture. When working with scales and measures that were developed without item response theory,

as the ESQ described here, it is common to find widely varying measurement precision levels for different values of the latent trait, which are difficult to summarize with any single index.

Scoring under the multidimensional IRT 'Correlated Traits' model

When every test item indicates one trait only (independent-clusters structure, as in Figure 2b), every trait may be considered separately for scoring purposes. Because each item response is conditional on one trait only, there is no difference between maximum likelihood trait score estimation using a correlated traits model or using separate unidimensional models, if the item parameters in both estimations are identical (which might not be the case).

There is, however, a difference between the trait-by-trait estimation and the multivariate estimation when Bayesian estimators are used. In a correlated traits model, relationships between the latent traits will alter the multivariate normal distribution used as a prior. For example, when two traits are positively correlated, combination of trait scores (θ_1=1, θ_2=1) is more likely than combination (θ_1=1, θ_2=–1). When the corresponding prior distribution is used in MAP or EAP estimation, one trait can "borrow strength" from related traits. The multivariate normal prior alters the posterior likelihood, and favors trait scores that are similar rather than different.

Multivariate priors add information and therefore enhance the measurement precision, which may be desired for shorter tests. However, their use has been criticized, notably by McDonald, who argued that they "corrupt measurement" in tests with independent clusters and correlated traits, because estimation is influenced by "indicators of conceptually distinct traits" (McDonald 2011, 531). Apart from philosophical concerns, the psychometric concern is that the multivariate prior makes errors of measurement correlated, even for independent clusters.

Illustration: ESQ satisfaction facets scored by the multidimensional EAP method

We estimate item parameters according to the correlated traits model (see Figure 2b), using the ML estimator. Loglikelihood and BIC values for this model are reported in Table 1. Because this model and the unidimensional models are nested, their relative goodness of fit can be compared using the likelihood ratio test. The difference between two loglikelihood values is multiplied by 2, and the resulting value (51.94) has a chi-square distribution with the degrees of freedom equal to the difference between the number of estimable parameters (37-36=1). As expected, the correlated trait model fits significantly better than the unidimensional model.

Table 2 gives the item slopes; Table 3 gives the item thresholds. The model-based correlation between Satisfaction with Care and Satisfaction with Environment is 0.59, which is very close to the correlation estimated in the exploratory factor model (0.60). The strong positive correlation ensures that the traits will "borrow strength" from each other when Bayesian estimation is used.

As the number of traits is small ($T = 2$) in this model, we can produce person scores using the EAP estimator easily. When more traits are measured, the MAP estimator would be more efficient. The EAP scores estimated using the correlated traits model range from −3.02 to 0.87 (mean = 0.003, SD = 0.87) for the Care facet; and they range from −2.84 to 0.86 (mean = 0.002, SD = 0.74) for the Environment facet. It can be seen that the standard deviations are much smaller than SD = 1 assumed for the latent trait. By assuming the standard multivariate prior, the EAP estimator shrank the scores for both scales. The shorter Environment scale is shrunken more severely.

INSERT FIGURE 4 ABOUT HERE

The standard errors associated with all EAP estimated trait scores in the sample are plotted in Figure 4. To compute the empirical reliability of the EAP scores, we use the squared standard errors of observed scores. For the Care scale, the standard errors range between 0.17 and 0.73. The squared values of standard errors average at 0.234, the variance of the observed score is 0.762, and the reliability is $\rho = 1-(0.234/0.762) = \mathbf{0.69}$. For comparison, Cronbach's alpha for the nine items forming the Care scale is impressive 0.93.

For the shorter Environment facet, the standard errors are much larger, ranging between 0.56 and 0.92. The squared values of standard errors average at 0.447, the variance of the observed score is only 0.553, and the reliability is $\rho = 1-(0.447/0.553) = \mathbf{0.19}$, unacceptable by any standards. For comparison, Cronbach's alpha for the three items forming the Environment scale is 0.49. Again, this example illustrates the danger of summarizing the conditional standard errors in IRT into a single coefficient, especially when the trait score is substantially shrunken due to Bayesian estimation, as is the case with the Environment facet.

Scoring Under the Multidimensional IRT Bifactor Model

A bifactor model assumes that item responses are caused by **two factors** (hence the name, *bi*-factor) – a general factor that influences all items, and one or more further specific factors – each influencing only a group of items. Specific factors are residuals left over after accounting for the general factor, and therefore represent specific, unique features common to a group of items that are not explained by the general factor. As all residuals, specific factors are assumed uncorrelated with the general factor and with each other.

Because any IRT bifactor model is a special case of the multidimensional IRT model given by (0.1), general formulae can be easily adopted to produce specialized item

information functions for the bifactor model. These specialized formulae are given in Appendix C.

Illustration: ESQ scored under the Bifactor IRT model

Here we fit a bifactor model to the ESQ data. In the model illustrated in Figure 2c, a common factor represents Affective Satisfaction influencing responses to all items. This factor explains all shared variance in the items describing the environment surrounding treatment (items 8-10); however, an additional specific factor is needed to capture the remaining shared variance in nine items describing care-related experiences. This specific factor cannot be thought of as 'Satisfaction with Care' because it captures common features of care-related items once the overall satisfaction has been accounted for (Brown et al. 2012). Rather, the specific factor could be named 'Experience of Care'.

We test this model, again using the ML estimator. The model estimates 12 slopes for the general factor, and 9 slopes for the specific factor. There are more estimable parameters in this model than in any other tested model, and not surprisingly, the loglikelihood reported in Table 1 is the largest. Because this model and the unidimensional models are nested, their relative goodness of fit can be compared using the likelihood ratio test. The resulting value ($\Delta\chi^2 = 87.64$), tested against the chi-square distribution with 45-36=9 degrees of freedom is highly significant. Therefore, the bifactor model fits significantly better than the unidimensional model. Comparing the bifactor model to the rival "correlated traits" model using the Bayesian information criteria, however, reveal that the more parsimonious correlated traits model may be preferable (its BIC is the smallest of all alternative models).

Table 2 gives the item slopes; Table 3 gives the item thresholds for the bifactor model. Using these parameters, we estimate the general and specific factor scores by the EAP method. Alternatively, the MAP estimator could be used here. The EAP scores range from –2.97 to 0.85 (mean = 0.00, SD = 0.75) for the general factor Affective Satisfaction; and they

range from –2.38 to 1.35 (mean = 0.00, SD = 0.75) for specific Experience with Care. It can be seen from the standard deviations (much smaller than the assumed standard deviation of 1 for the latent trait) that the EAP estimator shrank the scores for both factors severely.

Once the person EAP scores have been estimated together with their standard errors, we can plot them against each other for every person *j*. Because the standard errors for both the general and the specific factors are conditional on **two** factor scores, 3-D scatter plots are more suitable than the 2-D scatters. Figure 5 shows plots of the standard errors fully conditioned on both the general and specific factor scores. It can be seen that the standard errors are highest for parents who experience high levels of affective satisfaction, as well as evaluate specific experiences with care highly. This means that the ESQ does not differentiate well between parents who are satisfied with the service they received. For moderate to low scores on either satisfaction or experience of care, the standard errors are lower, reaching the minimum of 0.46 for the general and 0.43 for the specific factors. For many parents, however, the scores are estimated with much lower precision than that.

INSERT FIGURE 5 ABOUT HERE

To compute the empirical reliability of the EAP sample scores, we use the squared standard errors of estimated factor scores. For Affective Satisfaction, the standard errors range between 0.46 and 0.86. The squared values of standard errors average at 0.432, the variance of the observed score is only 0.570, and the empirical reliability is unacceptably low $\rho = 1–(0.432/0.570) =$ **0.24**. For Experience with Care, the standard errors range between 0.43 and 0.88. The squared values of standard errors average at 0.442, the variance of the observed score is only 0.557, and the empirical reliability is also very low $\rho =$

1–(0.442/0.557) = **0.21**. The resulting empirical reliability likely misrepresents the measurement precision for most of the latent trait due to the severe shrinkage of the EAP scores (which would also be true for MAP scores). Again, this example illustrates the danger of summarizing the conditional standard errors in IRT into a single coefficient, especially when the trait score is shrunken due to Bayesian estimation.

Evaluation of the alternative scoring methods for ESQ data example

In this empirical example, we illustrated a range of measurement models that can be applied to the ESQ to produce scores on slightly different conceptual constructs (i.e. global satisfaction; facets of satisfaction; or affective satisfaction (halo) separated from specific aspects of care). We illustrated the EAP scoring under all these models using the same dataset. We concluded that the unidimensional model is deficient in that it does not reflect the 2-factor structure that underlies these data. The other models, which address the multidimensional structure of the ESQ, fit the observed data significantly better.

Which model is the best to adopt when scoring the ESQ? Apart from the purpose of measurement, this depends on model properties, specifically: 1) what constructs the model measures; and, 2) how accurately these constructs are measured.

Measured constructs in the three alternative models

The unidimensional model assumes that the ESQ measures just one trait, which we tentatively named Satisfaction with Service. This scoring model would reflect the default approach to scoring the ESQ without investigating its factorial structure; the summated score would be the closest classical counterpart to the IRT score derived from this model. Does the measured construct actually represent Satisfaction with Service (i.e. overall satisfaction with all experiences)? Looking at correlations between this score and scores from the other models in Table 5, it becomes clear that the unidimensional model yields a common factor that is almost identical to Satisfaction with Care construct from the correlated traits model (r =

.999). Thus, it appears that the construct measured by the unidimensional model is essentially Satisfaction with Care. Nine items contribute strongly to the measurement of this construct; the remaining three items provide almost no contribution – their weak positive loadings on the common factor merely reflect the overall halo effect.

INSERT TABLE 5 ABOUT HERE

Examining correlations between the scores estimated under the alternative models given in Table 5 further, it could be seen that Affective Satisfaction (general factor assessed by the bifactor model) is closest in meaning to Satisfaction with Environment from the correlated traits model (r = .945). This implies that these two constructs capture all non-specific aspects of satisfaction – affect that colours parents' perceptions of their experiences. It also implies that Satisfaction with Environment construct probably has little to do with the environment surrounding treatment; rather, it captures aspects of satisfaction not related to Care.

Finally, Satisfaction with Care assessed by the correlated traits model is strongly related to the residual factor of the bifactor model – Experience of Care (r = .858). These constructs are, however, not the same. While Satisfaction with Care includes the affective element (hence its strong positive relationship with Satisfaction with Environment, r = .756), Experience of Care describes common features of care-related experiences controlling for the affective element of satisfaction.

Measurement precision provided by the three alternative scoring models

The marginal reliability coefficients painted quite a mixed picture of measurement precision provided by the ESQ. Empirical reliabilities are consistently lower than Cronbach's alpha for the same scales, and sometimes are unacceptably low. Considering challenging

features of this instrument – its profound ceiling effect and poor measurement accuracy at the top end of the latent trait, and the substantial shrinkage of scores by the use of Bayesian estimation – the marginal reliability does not reflect the measurement precision of individual trait scores. Indeed, looking at the range of standard errors for the Satisfaction with Care construct depicted in Figure 4, it can be seen that in the range from $\theta = -3$ to $\theta = 0$, the standard errors are small indeed. Many scores in this region have associated standard errors of around 0.2, and no scores have standard errors above 0.5. Using formula (0.19), the theoretical reliability corresponding to SE = 0.2 is .96; and the reliability corresponding to SE = 0.5 is .75. However, the above-average scoring half of the sample had large standard errors (between from about 0.5 to 0.73), corresponding to reliabilities from 0.75 to 0.47. The overall figure for the empirical reliability of 0.69 provides merely an aggregated picture. The aggregated picture would be more representative in applications where the standard error function is uniform, the estimated scores are distributed approximately normally, and the shrinkage is small.

Which measurement model to choose?

Which model is the most suitable for a particular instrument should be a decision based on conceptual, statistical and practical grounds. A particular measurement focus (e.g. whether the general factor or the facet factors are of interest) imposes practical requirements on the scoring model. Approaching the hypothesized model choice from the theoretical perspective governing the instrument's design, the Experience of Service Questionnaire used as an example in this chapter was constructed to measure one construct – satisfaction with service. The unidimensional model is clearly suitable from this point of view. However, this approach "masks" the near-zero contribution of Environment-related items, and the real meaning of the measured construct, which is nearly identical to care-related satisfaction (Satisfaction with Care).

The correlated traits model, on the other hand, fits the data significantly better and enables measurement of two facets of satisfaction. The Satisfaction with Care scale representing care-related aspects of satisfaction is almost as reliable as the total scale. The Satisfaction with Environment scale representing non-specific aspects of satisfaction is unreliable and is not useful for any practical purposes.

The bifactor model yields the best fit to these data, and enables measurement of two independent factors influencing item responses – Affective Satisfaction and Experience of Care. In our view, this model provides the most theoretically sound picture of these data. The model separates two sources of variance, and, unlike the other alternative models, provides an adequate explanation of strong dependencies between experiences that are theoretically unrelated for parents attending one service (i.e. facilities, location, appointment times). These dependencies are fully explained by the general halo or 'affective overtones' factor, while care-specific experiences are explained by a separate factor. However, both constructs suffer from low accuracy of estimation for significant parts of the latent traits (specifically above average satisfaction and experience of care). This makes the bifactor measurement model less useful for scoring the ESQ in practice.

Acknowledgements

The author is grateful to the CORC collaboration for providing data for the empirical example used in this chapter, and to the CAMHS Outcomes Research Consortium (CORC) central team researchers Jenna Bradley and Halina Flannery for preparing the data for analyses.

References

Ackerman, Terry A. 2005. “Multidimensional item response theory modeling”. In *Contemporary Psychometrics*, edited *by* Albert Maydeu-Olivares and John J. McArdle, 3-26. Mahwah, NJ: Lawrence Erlbaum.

Attride-Stirling, Jennifer. 2002. *Development of methods to capture users’ views of CAMHS in clinical governance reviews.* http://www.corc.uk.net/resources/downloads/

Bock, R. Darrell. 1975. “*Multivariate statistical methods in behavioral research”*. New York: McGraw-Hill.

Brown, Anna, Tamsin Ford, Jessica Deighton, and Miranda Wolpert. 2012. "Satisfaction in child and adolescent mental health services: Translating users’ feedback into measurement." *Administration and Policy in Mental Health and Mental Health Services Research*: 1-13.

Brown, Anna, and Alberto Maydeu-Olivares. 2011. “Item response modeling of forced-choice questionnaires”. *Educational and Psychological Measurement,* 71(3), 460-502.

Dodd, Barbara G., R. J. De Ayala, and William R. Koch. 1995. "Computerized adaptive testing with polytomous items." *Applied psychological measurement* 19, no. 1: 5-22.

Du Toit, Matilda. 2003. “*IRT from Sscientific Software International”*. Chicago: Scientific Software International.

Embretson, Susan E., and Steven P. Reise. 2000. “*Item response theory”*. Mahwah, NJ: Erlbaum Publishers.

Green, Bert F., R. Darrell Bock, Lloyd G. Humphreys, Robert L. Linn, and Mark D. Reckase. 1984. "Technical guidelines for assessing computerized adaptive tests." *Journal of Educational Measurement*, 21, no. 4: 347-360.

Gibbons, Robert D., Jason C. Immekus, and R. Darrell Bock. 2007. “The Added Value of Multidimensional IRT Models. Didactic workbook”. Accessed June 1 http://outcomes.cancer.gov/areas/measurement/multidimensional_irt_models.pdf

Masters, Geofferey N., and Benjamin D. Wright. 1997. "The partial credit model." In *Handbook of modern item response theory,* edited by Wim J. van der Linden and Ronald Hambleton, 101-121. New York: Springer.

McDonald, Roderick P. 1999. “*Test theory. A unified approach”*. Mahwah, NJ: Lawrence Erlbaum.

McDonald, Roderick P. 2011. “Measuring latent quantities”. *Psychometrika, 76:* 511-536.

Muthén, Linda K., and Bengt O. Muthén. 1998-2012. “*Mplus User’s guide. Seventh edition”*. Los Angeles, CA: Muthén & Muthén.

Reckase, Mark. 2009. “*Multidimensional Item Response Theory”*. New York: Springer.

Reise, Stephen, and Mark Haviland. 2005. “Item response theory and the measurement of clinical change”. *Journal of Personality Assessment,* 84(3): 228-238.

Samejima, Fumiko. 1969. “*Estimation of Latent Ability Using a Response Pattern of Graded Scores”* (Psychometric Monograph No. 17). Richmond, VA: Psychometric Society. Accessed July 1 http://www.psychometrika.org/journal/online/MN17.pdf

Samejima, Fumiko. 1997. “Graded response model”. In *Handbook of modern item response theory*, edited by Wim J. van der Linden and Ronald Hambleton, 85-100. New York: Springer.

Thissen, David, Lauren Nelson, and Kimberly A. Swygert. 2001. “Item response theory applied to combinations of multiple-choice and constructed-response items – approximation methods for scale scores”. In *Test Scoring*, edited by David Thissen and Howard Wainer, 179-216. Mahwah, NJ: Lawrence Erlbaum.

Thissen, David, and Maria Orlando. 2001. “Item response theory for items scored in two categories”. In *Test Scoring*, edited by David Thissen and Howard Wainer, 73-140. Mahwah, NJ: Lawrence Erlbaum.

van der Linden, Wim J. and Ronald Hambleton, R. 1997. “Item Response Theory: brief history, common models and extensions”. In *Handbook of modern item response theory*, edited by Wim J. van der Linden and Ronald Hambleton, 1-28. New York: Springer.

Appendix A. Computation of EAP scores and their Standard Errors

The **EAP score** for one trait is computed as the ratio between the integral of the posterior function weighted by the latent trait θ, and the unweighted integral of the posterior function. The EAP scores are approximated using numerical integration of the posterior distribution along the latent trait continuum as follows:

$$\mathrm{EAP}\left(\mathbf{u}\middle|\theta\right) \approx \frac{\sum_q l\left(\mathrm{u}_1,\mathrm{u}_2,...,\mathrm{u}_m\middle|\theta_q\right)\phi\left(\theta_q\right)\theta_q d\theta_q}{\sum_q l\left(\mathrm{u}_1,\mathrm{u}_2,...,\mathrm{u}_m\middle|\theta_q\right)\phi\left(\theta_q\right)d\theta_q}. \tag{0.23}$$

In this expression, $l\left(\mathrm{u}_1,\mathrm{u}_2,...,\mathrm{u}_m\middle|\theta_q\right)$ denotes the likelihood of the observed response pattern defined in (0.2), $\phi(x)$ is the standard normal density function evaluated at x (here, at each of the quadrature points); and $d\theta_q$ denotes the size of the interval between two adjacent quadrature points. The standard error of the EAP score is computed at the same quadrature points as follows:

$$\mathrm{SE}_{EAP} \approx \sqrt{\frac{\sum_q l\left(\mathrm{u}_1,\mathrm{u}_2,...,\mathrm{u}_m\middle|\theta_q\right)\phi\left(\theta_q\right)\left(\theta_q-\mathrm{EAP}\right)^2 d\theta_q}{\sum_q l\left(\mathrm{u}_1,\mathrm{u}_2,...,\mathrm{u}_m\middle|\theta_q\right)\phi\left(\theta_q\right)d\theta_q}} \tag{0.24}$$

Appendix B. Experience of Service Questionnaire (parent version)

Response options: *Certainly True – Partly True – Not True – (Don't know)*

("don't know" response option is considered missing data and is not scored).

1. I feel that the people who have seen my child listened to me
2. It was easy to talk to the people who have seen my child
3. I was treated well by the people who have seen my child
4. My views and worries were taken seriously
5. I feel the people here know how to help with the problem I came for
6. I have been given enough explanation about the help available here
7. I feel that the people who have seen my child are working together to help with the problem(s)
8. The facilities here are comfortable (e.g. waiting area)
9. The appointments are usually at a convenient time (e.g. don't interfere with work, school)
10. It is quite easy to get to the place where the appointments are
11. If a friend needed similar help, I would recommend that he or she come here
12. Overall, the help I have received here is good

Appendix C. Item Information Function for a bifactor model

Here we give formulae necessary to compute the item and test information for the bifactor model. Let *g* be a **general factor** measured by a test, and $s_1, s_2, \ldots, s_T$ be *T* **specific factors**. The set of factors underlying item responses is therefore $\mathbf{g}^* = (g, s_1, s_2, ..., s_T)'$. The general factor and the specific factors are assumed uncorrelated with each other, and have zero means and unit variances so that their distribution is multivariate standard normal $\mathbf{g}^* \sim N_{T+1}(\mathbf{0}, \mathbf{I})$, where **I** denotes the identity covariance matrix. Under this model, the response to item *i* is influenced by two factors – the general factor *g* and one specific factor, say s_a

$$P_i(\mathbf{g}^*) = P(\mathrm{u}_i = 1 | \mathbf{g}^*) = \Phi(-\tau_i + \beta_{0i} g + \beta_{ai} \mathrm{s}_a), \tag{0.25}$$

where, τ_i is the threshold, β_{0i} is the slope for the general factor *g*, and β_{ai} is the slope for the specific factor s_a.

Because the general factor and the specific factors in this model are orthogonal, the ML item information **about the general factor** *g* is computed substituting the gradient in equation (0.11) with:

$$\nabla_g P_i(\mathbf{g}^*) = \frac{\partial P_i(\mathbf{g}^*)}{\partial g}. \tag{0.26}$$

Now, the partial derivative with respect to *g* is

$$\frac{\partial P_i(\mathbf{g}^*)}{\partial g} = \beta_{0i} \phi(-\tau_i + \beta_{0i} g + \beta_{ai} \mathrm{s}_a), \tag{0.27}$$

where $\phi(z)$ is the normal density function evaluated at *z* (McDonald 1999; page 284). Thus, the IIF in direction of the general factor *g* is given by

$$\mathcal{I}_i^g\left(\mathbf{g}^*\right)=\frac{\left(\beta_{0i}\right)^2\left[\phi\left(-\tau_i+\beta_{0i}g+\beta_{ai}s_a\right)\right]^2}{\Phi\left(-\tau_i+\beta_{0i}g+\beta_{ai}s_a\right)\left[1-\Phi\left(-\tau_i+\beta_{0i}g+\beta_{ai}s_a\right)\right]}. \tag{0.28}$$

Note that the item information function for the bifactor model is fully conditioned on the general and the specific factors, therefore local independence holds and the item information functions are additive. The test information about the general factor *g* is the sum of all IIFs in the direction of *g*. This summation, however, will make the test information function for the general factor **conditional on all specific factors**, although each item information function is only conditional on **one** specific factor (in addition to the general factor).

When Bayesian estimation is used, the prior information must be added to the ML test information to compute the posterior test information. In bifactor models, the latent covariance matrix is the identity matrix, $\mathbf{\Sigma} = \mathbf{I}$, and the amount of information added by the multivariate normal prior is simply 1,

$$\mathcal{I}_P^g\ \mathbf{g}^*\ =\mathcal{I}^g\ \mathbf{g}^*\ +\left[\mathbf{\Sigma}^{-1}\right]_g=\sum_i\mathcal{I}_i^g\ \mathbf{g}^*\ +1. \tag{0.29}$$

Tables

Table 1. Goodness of fit for the three alternative ESQ models

Model	Loglikelihood	Number of parameters	BIC
Unidimensional	-3989.27	36	8215.18
Correlated traits	-3963.29	37	8169.81
Bifactor	-3945.45	45	8186.71

Note. BIC = Bayesian Information Criterion.

Table 2. Slopes for the three alternative ESQ models

		Unidimensional	Correlated traits		Bifactor	
Item		Factor 1 Satisfaction	Factor 1 Care	Factor 2 Environment	Factor 1 General	Factor 2 Specific
1	listened	2.04	2.04		1.57	1.39
2	easy to talk	1.54	1.54		1.37	0.98
3	treated well	1.77	1.76		2.23	1.19
4	taken seriously	2.29	2.30		1.48	1.72
5	know how to help	2.09	2.11		1.23	1.83
6	given explanation	1.60	1.61		0.98	1.28
7	working together	2.26	2.27		1.36	1.94
8	comfortable facilities	0.43		0.74	0.65	
9	convenient times	0.46		0.97	0.76	
10	convenient location	0.40		0.69	0.67	
11	recommend to a friend	2.55	2.53		1.75	1.85
12	good help overall	3.18	3.24		2.13	3.11

Note. All slopes are significant at the .001 level.

Table 3. Thresholds for the three alternative ESQ models

		Unidimensional		Correlated traits		Bifactor	
Item		Thresh.1	Thresh.2	Thresh.1	Thresh.2	Thresh.1	Thresh.2
1	listened	-4.36	-2.17	-4.35	-2.16	-4.46	-2.22
2	easy to talk	-3.49	-1.88	-3.48	-1.87	-3.73	-1.99
3	treated well	-4.79	-2.73	-4.74	-2.71	-6.44	-3.66
4	taken seriously	-4.07	-2.21	-4.07	-2.21	-4.04	-2.19
5	know how to help	-3.07	-0.93	-3.08	-0.93	-3.20	-0.97
6	given explanation	-2.28	-0.78	-2.28	-0.77	-2.30	-0.78
7	working together	-3.61	-1.64	-3.60	-1.63	-3.72	-1.69
8	comfortable facilities	-2.36	-1.02	-2.69	-1.16	-2.58	-1.12
9	convenient times	-1.18	-0.25	-1.49	-0.31	-1.36	-0.29
10	convenient location	-1.56	-0.68	-1.76	-0.76	-1.75	-0.76
11	recommend to a friend	-3.90	-2.37	-3.86	-2.34	-3.89	-2.37
12	good help overall	-5.10	-2.29	-5.15	-2.30	-5.91	-2.66

Table 4. Reliability summary for the three alternative ESQ models

Model and measured construct	Min SE	Max SE	Empirical reliability	Alpha*
Unidimensional				
Satisfaction with Service	0.17	0.73	.70	.90
Correlated traits				
Satisfaction with Care	0.17	0.73	.69	.93
Satisfaction with Environment	0.56	0.92	.19	.49
Bifactor				
Affective Satisfaction	0.46	0.86	.24	--
Experience of Care	0.43	0.88	.21	--

Note. Cronbach's alpha is calculated assuming that the item responses are continuous; it is not model-based and is given here for comparison only.

Table 5. Correlations between ESQ global satisfaction scores estimated using the three alternative scoring models

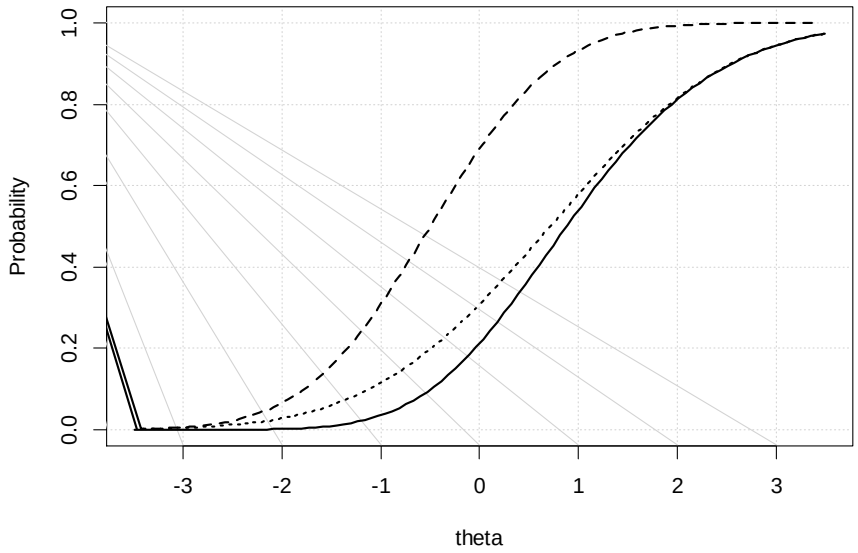
	Correlated traits		Bifactor	
Model and measured construct	S.Care	S.Env.	Aff.Sat.	Exp.Care
Unidimensional				
Satisfaction with Service	.999	.777	.814	.842
Correlated traits				
Satisfaction with Care		.756	.796	.858
Satisfaction with Environment			.945	.362
Bifactor				
Affective Satisfaction				.373
Experience of Care				

Note. S.Care = Satisfacion with Care; S.Env. = Satisfaction with Environment; Aff.Sat. = Affective Satisfaction; Exp.Care = Experience of Care.

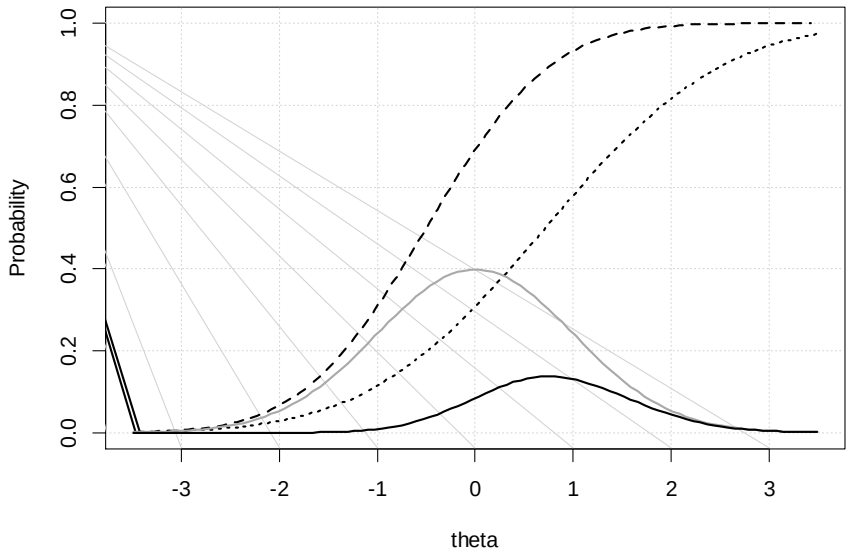
Figures

Figure 1. Example likelihood functions for two endorsed items ($u_1 = 1$, $u_2 = 1$)

(a) Likelihood based on item responses only



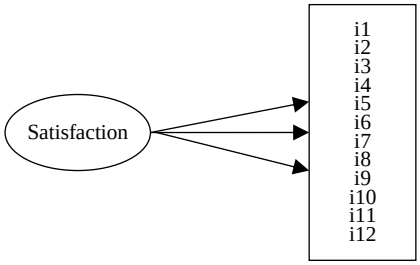
(b) Likelihood based on item responses and population distribution (standard normal)



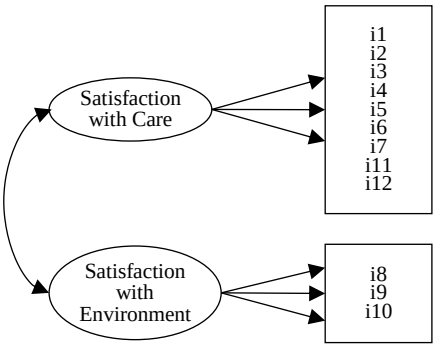
Note. The dashed line is the IRF for item 1; the dotted line is the IRF for item 2; the solid grey line is the normal density function; the solid black line in (a) is the total likelihood function, and in (b) it is the posterior likelihood function.

Figure 2. Three alternative models for ESQ item responses

a. Unidimensional model



b. Correlated Traits model



c. Bifactor model

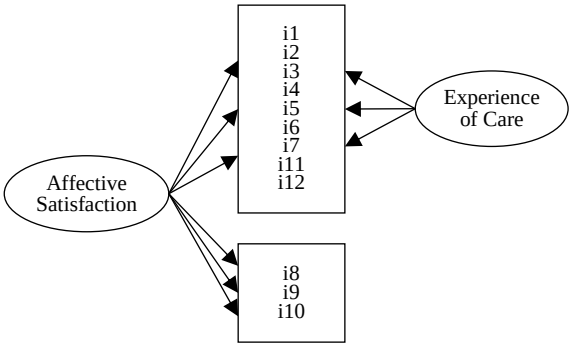


Figure 3. Standard Errors of the EAP Satisfaction scores under the unidimensional model

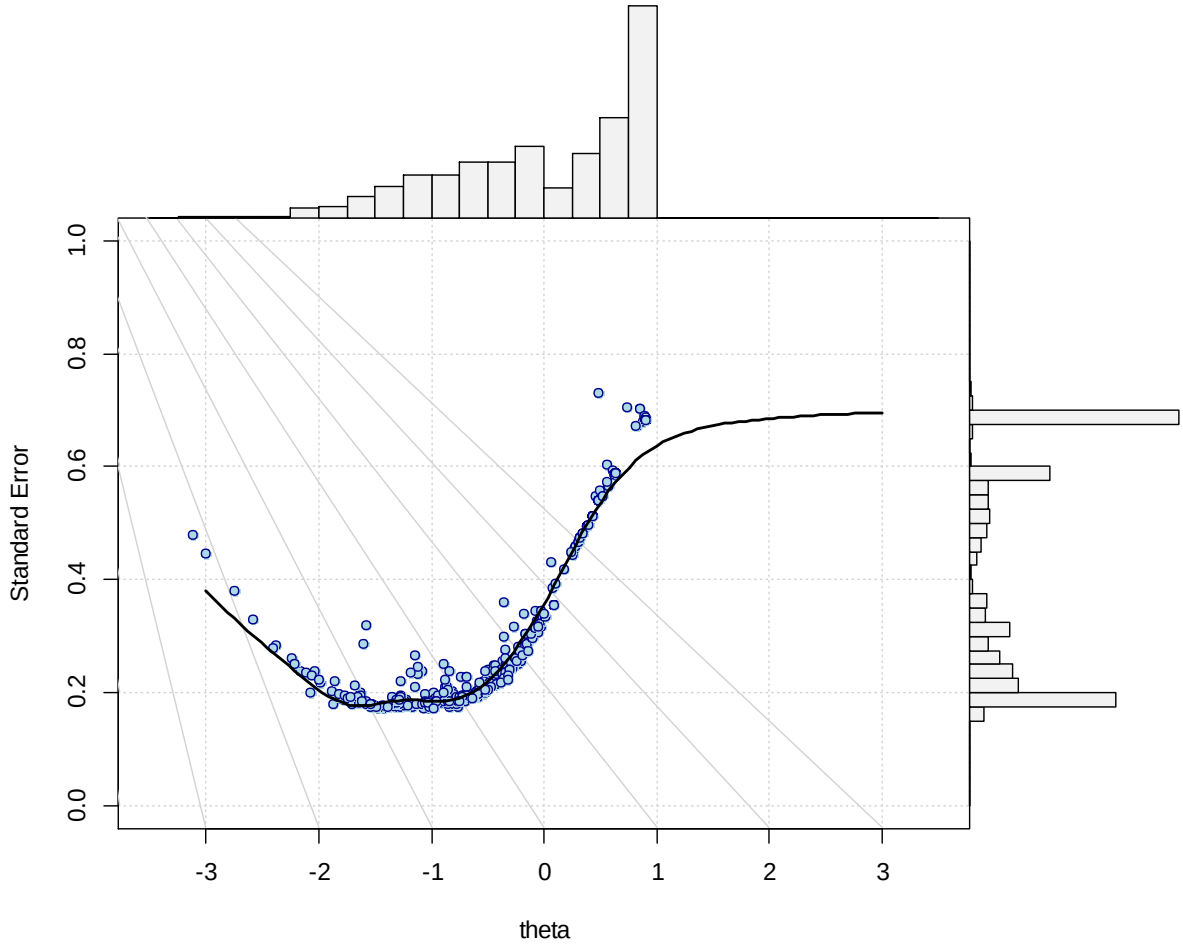
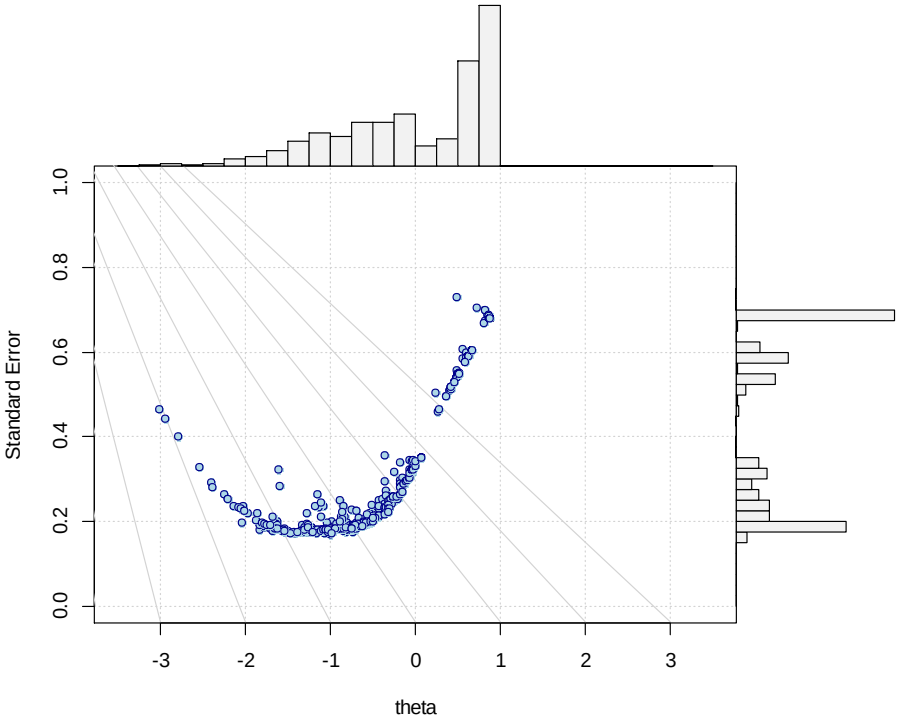


Figure 4. Standard Errors of the EAP scores under the correlated traits model

(a) Satisfaction with Care



(b) Satisfaction with Environment

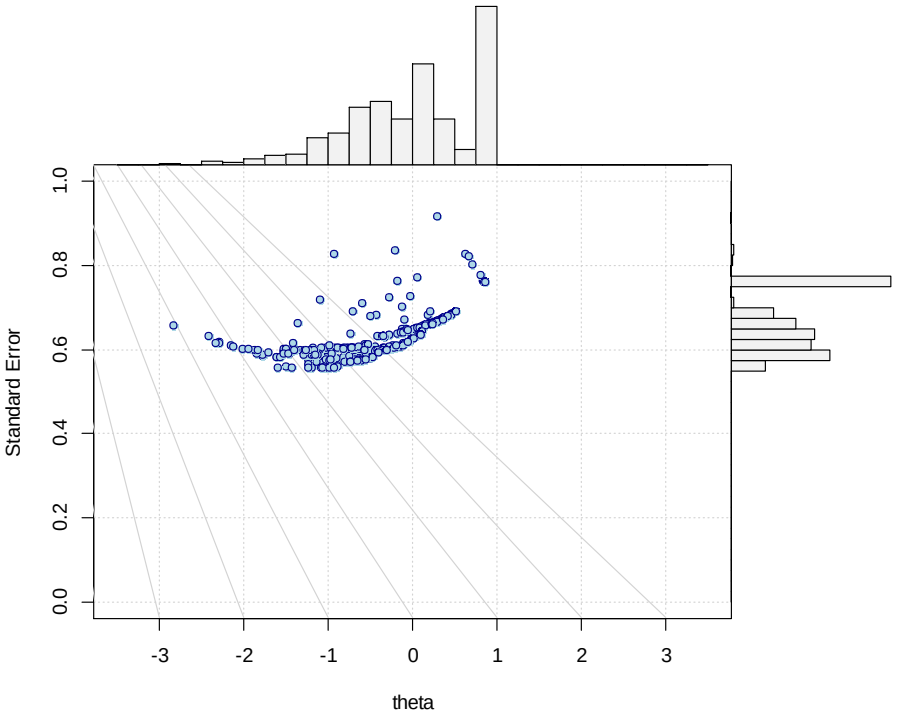
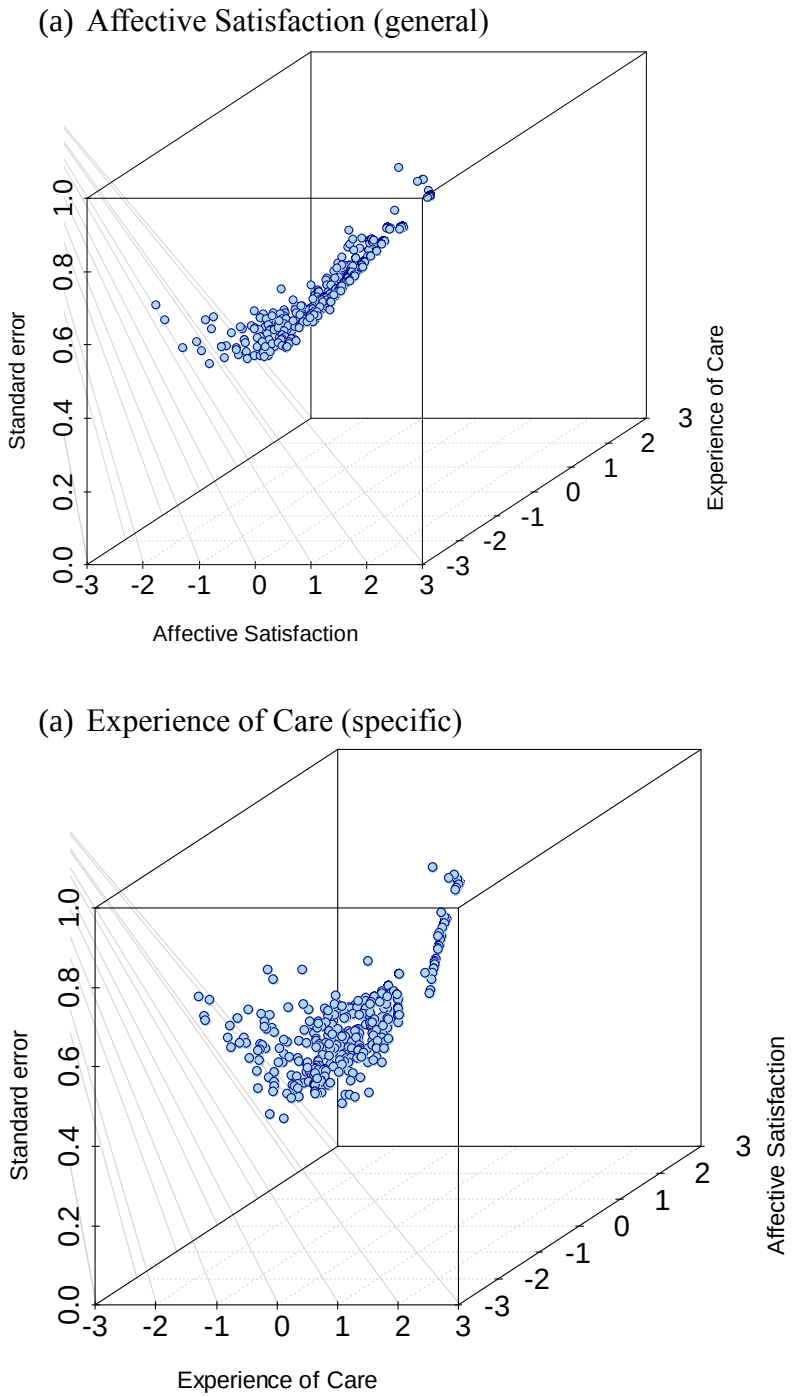


Figure 5. Standard Errors of the EAP scores under the bifactor model



[i] This parameterization is convenient with multidimensional IRT models. With unidimensional models, an alternative IRT parameterization is often used, whereby the item *discrimination* a_i and *difficulty* b_i are defined so that $P_i = \Phi(a_i[\theta - b_i])$. Thus, the discrimination is equivalent to the slope, and the difficulty equals the threshold divided by the slope.

[ii] This is true for all item response functions where the probability of a keyed response is monotonously increasing or decreasing, a so-called dominance response process.

[iii] When we use the logistic link function $L(x)$, the information function amounts to

$$\mathcal{I}_i(\theta) = \left(D\beta_i\right)^2 L\left(-D\tau_i + D\beta_i\theta\right)\left[1 - L\left(-D\tau_i + D\beta_i\theta\right)\right].$$

All formulae for item information involving the normal-ogive link function given further in this chapter can be adopted for the logistic link by using this expression. D=1.7 is the scaling constant used with the normal ogive item parameters